Stochastic Intrinsic Kriging for Simulation Metamodelling

Mehdad, Ehsan; Kleijnen, J.P.C.

Publication date:
2015

Document Version
Early version, also known as pre-print

No. 2015-038

STOCHASTIC INTRINSIC KRIGING FOR SIMULATION METAMODELLING

By

Ehsan Mehdad, Jack P.C. Kleijnen

22 July, 2015

This is a revised version of CentER Discussion Paper

No. 2014-054

ISSN 0924-7815
ISSN 2213-9532

Stochastic Intrinsic Kriging for Simulation Metamodelling

Ehsan Mehdad and Jack P.C. Kleijnen
Tilburg School of Economics and Management, Tilburg University

July 21, 2015

Abstract

Kriging provides metamodels for deterministic and random simulation models. Actually, there are several types of Kriging; the classic type is so-called *universal Kriging*, which includes *ordinary Kriging*. These classic types require estimation of the *trend* in the input-output data of the underlying simulation model; this estimation deteriorates the Kriging metamodel. We therefore consider so-called *intrinsic Kriging* originating in geostatistics, and derive intrinsic Kriging for deterministic and random simulations. Moreover, for random simulations we derive experimental designs that specify the number of replications that varies with the input combination of the simulation model. To compare the performance of intrinsic Kriging and classic Kriging, we use several numerical experiments with deterministic simulations and random simulations. These experiments show that intrinsic Kriging gives better metamodels, in most experiments.

Keywords: simulation, Gaussian process, Kriging, intrinsic random functions, metamodel

JEL: C0, C1, C9, C15, C44

1 Introduction

Simulation is a popular scientific method for analyzing complex systems. However, simulation models often require much computer time to run. We may use the input/output (I/O) data of a relatively small experiment with the simulation model, to build a metamodel (also called an *emulator* or a *surrogate*). This metamodel gives an explicit simple approximation of the I/O function that is implicitly defined by the underlying simulation model. Next, we may use this metamodel for *sensitivity analysis* and *optimization*.

Though there are many types of metamodels (e.g., linear regression models), we focus on *Kriging* or *Gaussian process* models. The basic idea of Kriging is to predict the value of the output (response) at a new input combination (point) on the basis of the values of the output that have already been observed at

old points; these values form the I/O data. An additional advantage of Kriging is that it not only predicts the new output but also quantifies the uncertainty of this prediction. Classic textbooks on Kriging are Cressie (1991), Stein (1999), Santner et al. (2003) and Chilès and Delfiner (2012).

More precisely, a Kriging metamodel $Y(\mathbf{x})$ with output Y and continuous (real valued) d-dimensional input $\mathbf{x} \in \mathbb{R}^d$ is the sum of two components; namely, *drift* and *residual*. The drift is a smooth deterministic function that captures the systematic pattern of the simulation I/O function. The residual is a zero-mean random function that captures the erratic behavior of the I/O data. We define the following three types of Kriging using different assumptions about the drift: (1) *simple Kriging* with a known drift (2) *ordinary Kriging* (OK) with a constant unknown drift (3) *universal Kriging* (UK) with an unknown drift that consists of known basis functions (e.g., low-order polynomials) with unknown coefficients. In case of simple Kriging, we assume that the residual is *second-order stationary*. For non-stationary I/O data, we may use UK, but the drift creates *bias* in the estimation of the covariance function of the residual; see Chilès and Delfiner (2012, p. 125).

A solution for this bias problem removes the drift by considering the *increments* $Y(\mathbf{x}) - Y(\mathbf{x}_0)$ where $\mathbf{x}_0$ is a given point. This solution gives rise to a broader category of random functions called *intrinsic random functions* (IRFs). An intrinsic stationary random function can be non-stationary but its increments are second-order stationary. Originally, the French mathematician George Matheron discussed and generalized the concept of IRFs for Kriging; see Matheron (1973). He not only formalized Kriging but also extended it to *intrinsic Kriging* (IK). To the best of our knowledge, IK has been investigated in simulation in Vazquez et al. (2005) only. However, Vazquez et al. (2005) considers only deterministic simulation, whereas we consider deterministic and random simulations. Vazquez et al. (2005) uses IK together with an additional set of factors to transform a black-box model into a grey-box one, whereas we use IK to study a *black-box* model; i.e., we assume that we observe only the I/O data resulting from the simulation experiment. Vazquez et al. (2005) considers a single covariance function, whereas we consider several covariance functions. Finally, Vazquez et al. (2005) does not compare IK with classic alternatives such as OK, whereas we compare IK and OK—and also their stochastic counterparts.

Therefore it seems interesting to formalize IK for a simulation readership, and to compare the performance of Kriging and IK in experiments with test functions. For all these experiments we use the root mean squared error (RMSE) as the performance criterion. In these experiments we study the effects of dimensionality, and the number of input combinations. The main conclusion will be that IK gives a more accurate metamodel than UK in deterministic simulation and random simulation.

We organize the rest of this paper as follows. Section 2 summarizes UK. Section 3 formalizes IRFs and their use in IK. Section 4 adapts IK for random simulation, including an experimental design for this type of simulation. Section 5 presents numerical experiments. Section 6 summarizes our conclusions.

2 Universal Kriging and Intrinsic Random Functions

The UK metamodel is

$$Y(\mathbf{x}) = \beta(\mathbf{x}) + M(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^d \tag{1}$$

where $\beta(\mathbf{x})$ is the *drift* and $M(\mathbf{x})$ is the *residual*. The drift is often assumed to be a low-order polynomial; i.e., $\beta(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$ where $\mathbf{f}(\mathbf{x})$ is a vector of $p+1$ known regression functions and $\boldsymbol{\beta}$ is a vector of $p+1$ unknown parameters.

In simple Kriging with a *known* drift, we assume $M(\mathbf{x})$ is a second-order stationary function with the covariance function $C(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(M(\mathbf{x}) - \beta)(M(\mathbf{x}') - \beta)]$. This covariance function is bounded; i.e., $|C(\mathbf{x}, \mathbf{x}')| \leq \tau^2$ where τ^2 is the variance of $M(\mathbf{x})$. In UK we do not know β so we cannot compute $M(\cdot) - \beta$ in $C(\mathbf{x}, \mathbf{x}')$. We therefore use another tool called the *structure function* or *variogram* $\gamma_M(\mathbf{x}, \mathbf{x}') = \text{Var}\left[M(\mathbf{x}) - M(\mathbf{x}')\right]/2$. The variogram shows the dissimilarity between $M(\mathbf{x})$ and $M(\mathbf{x}')$ for any pair of observed points. The variogram is bounded for a second-order stationary function, and increases to infinity for a non-stationary function. A second-order stationary function implies $\gamma_M(\mathbf{x}, \mathbf{x}') = \tau^2 - C_M(\mathbf{x}, \mathbf{x}')$. Unfortunately the drift affects γ_M. Chilès and Delfiner (2012, p. 125) shows that even with an *optimal* estimator of the drift, the variogram of the *estimated* residual is biased downward.

$$\gamma_M(\mathbf{x}, \mathbf{x}') = \gamma_Y(\mathbf{x}, \mathbf{x}') - \frac{1}{2}\text{Var}\left(\hat{\beta}(\mathbf{x}) - \hat{\beta}(\mathbf{x}')\right)$$

How should we then remove the drift from I/O data? We start with the simple case of a *constant drift*. To remove a constant drift, we use intrinsic random functions whose *increments* $Y(\mathbf{x}) - Y(\mathbf{x}_0)$ remove the constant drift; moreover, these increments are second-order stationary. This stationarity enables us to calculate the variance for the linear combination of increments. An IRF itself may not have a finite variance, but we can calculate the variance for 'allowable linear combinations' of IRF, defined as follows.

Definition 1. *If $Y(\mathbf{x})$ is an IRF, then any linear combination $\sum_i \lambda_i Y(\mathbf{x}_i)$ satisfying $\sum_i \lambda_i = 0$ is an 'allowable linear combination'.*

A linear combination of increments for an IRF is also an allowable linear combination since the coefficients of the two terms in each increment are ± 1. Conversely, any linear combination with $\sum_i \lambda_i = 0$ is equivalent to $\sum_i \lambda_i (Y(\mathbf{x}_i) - Y(\mathbf{x}_0))$ for any choice of $\mathbf{x}_0$. We can show that the covariance of two allowable linear combinations $\sum_i \lambda_i Y(\mathbf{x}_i)$ and $\sum_j \mu_j Y(\mathbf{x}_j)$ is

$$\text{Cov}\left(\sum_i \lambda_i Y(\mathbf{x}_i), \sum_j \mu_j Y(\mathbf{x}_j)\right) = -\sum_i \sum_j \lambda_i \mu_j \gamma_Y(\mathbf{x}_j - \mathbf{x}_i).$$

Allowable linear combinations of IRFs remove a constant drift and have finite variance. It is also possible to remove a more general drift (linear, quadratic, etc.) from the I/O data. We discuss this generalization in the next section.

3 Intrinsic random functions of order k, and IK

In the previous section, we showed that we can define a class of non-stationary functions called IRFs using allowable linear combination, which are equivalent to linear combinations of increments. *Ordinary* (zero-order) increments of an IRF are second-order stationary, they also eliminate a constant drift. In this section, we discuss a wider class of non-stationary functions called *intrinsic random function of order k* (IRF-k) whose increments of order k are second-order stationary, and also eliminate a polynomial drift of order k. To formalize the IRF-k, we follow Cressie (1991, pp. 299–306) and Chilès and Delfiner (2012, pp. 252–257). So, we rewrite (1) as

$$\boldsymbol{\mathscr{Y}} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\mathscr{M}} \tag{2}$$

where $\boldsymbol{\mathscr{Y}} = (\mathscr{Y}(\mathbf{x}_1), \ldots, \mathscr{Y}(\mathbf{x}_m))^\top$, and $\boldsymbol{\mathscr{M}} = (\mathscr{M}(\mathbf{x}_1), \ldots, \mathscr{M}(\mathbf{x}_m))^\top$. Let $\mathbf{Q}$ be an $m \times m$ matrix such that $\mathbf{QF} = \mathbf{O}$ where $\mathbf{O}$ is an $m \times (p+1)$ matrix with all elements zero. Together, $\mathbf{Q}$ and (2) give

$$\mathbf{Q}\boldsymbol{\mathscr{Y}} = \mathbf{Q}\boldsymbol{\mathscr{M}}.$$

Consequently, the second-order properties of $\mathbf{Q}\boldsymbol{\mathscr{Y}}$ depend on $\mathbf{Q}\boldsymbol{\mathscr{M}}$ and *not on the regression function* $\mathbf{F}\boldsymbol{\beta}$.

To generalize the model in (1), we need a stochastic process for which $\mathbf{Q}\boldsymbol{\mathscr{M}}$ is second-order stationary; such processes are called *intrinsically* stationary processes. We assume that $f_j(\mathbf{x})$ $(j = 1, \ldots, p+1)$ are mixed monomials $x_1^{i_1} \cdots x_d^{i_d}$ with $\mathbf{x} = (x_1, \ldots, x_d)^\top$ and nonnegative integers $i_1, \ldots, i_d$ such that $i_1 + \ldots + i_d \leq k$ with k a given nonnegative integer. An IRF-k is a random process $\mathscr{Y}$ for which $\sum_{i=1}^m \lambda_i \mathscr{Y}(\mathbf{x}_i)$ with $\mathbf{x}_i \in \mathbb{R}^d$ is an *allowable linear combination of order k* and second-order stationary. This $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ is called the *generalized-increment* vector which must satisfy the following conditions

$$(f_j(\mathbf{x}_1), \ldots, f_j(\mathbf{x}_m))\, \boldsymbol{\lambda} = 0 \ (j = 1, \ldots, p+1).$$

Let Λ_k be the class of all generalized-increments of order k. For a one-dimensional function, the generalized-increment vector $\boldsymbol{\lambda} \in \Lambda_k$ must satisfy the conditions

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i = 0, && \text{(constant drift)} \\
&\sum_{i=1}^m \lambda_i = 0, \ \sum_{i=1}^m \lambda_i x_i = 0, && \text{(linear drift)} \\
&\sum_{i=1}^m \lambda_i = 0, \ \sum_{i=1}^m \lambda_i x_i = 0, \ \sum_{i=1}^m \lambda_i x_i^2 = 0 && \text{(quadratic drift).}
\end{aligned}$$

It is clear that an IRF-k is also an IRF-$(k+1)$; i.e, $\Lambda_{k+1} \subset \Lambda_k$.

We give the following four IRF-k examples, following Matheron (1973) and Chilès and Delfiner (2012).

1. The $(k+1)$ integral of a zero-mean stationary random function $Z(t)$ is an IRF-k for $k = 0, 1, 2, \ldots$:

$$Z_k(x) = \int_0^x \frac{(x-t)^k}{k!} Z(t)dt.$$

   For the case $d = 1$ and $k = 0$ we have $Z_0(x) = \int_0^x Z(t)dt$. For any h the increment $Z_0(x+h) - Z_0(x) = \int_x^{x+h} Z(t)dt$ is stationary since it is the moving average of a stationary function $Z(t)$. The proof for the general k can be obtained through induction.

2. The same argument can be applied to an IRF-0; e.g., a Brownian motion. Integrating k times a Brownian motion $B(x)$ gives an IRF-k:

$$B_k(x) = \int_0^x \frac{(x-t)^k}{k!} B(t)dt.$$

3. An ARIMA process (autoregressive integrated moving average process) is a process whose finite differences of order k is a stationary ARMA process, so an ARIMA process is an IRF-$(k-1)$.

4. If $Z(x)$ is a random function which is differentiable $(k+1)$ times and if all its derivatives of order $(k+1)$ are stationary with zero mean, then $Z(x)$ is an IRF-k. This example characterizes a differentiable IRF-k; of course there are also non-differentiable IRF-k's.

We know that any linear combination of a second-order stationary function has a finite variance. In Section 2, we mentioned that an IRF itself does not have a finite variance or its variance may depend on $\mathbf{x}$. However, we can calculate the variance of a linear combination of the ordinary increments of an IRF (equivalent to allowable linear combination of order 0) in terms of a *variogram*. The covariance structure for any two allowable linear combinations of order k is called a *generalized covariance function* (GCF) K.

Now we discuss properties of K. We follow Matheron (1973), Cressie (1991, pp. 304–305), and Chilès and Delfiner (2012, pp. 257–269). Obviously K is symmetric; so $K(\mathbf{x}_i, \mathbf{x}_{i'}) = K(\mathbf{x}_{i'}, \mathbf{x}_i)$, and K must be *conditionally* positive definite so

$$\text{var}(\boldsymbol{\lambda}^\top \mathscr{Y}) = \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} K(\mathbf{x}_i - \mathbf{x}_{i'}) \geq 0, \forall \boldsymbol{\lambda} \text{ such that } (f_j(\mathbf{x}_1), \ldots, f_j(\mathbf{x}_m))\, \boldsymbol{\lambda} = 0$$

where the condition must hold for $j = 1, \ldots, p+1$. This condition makes $\boldsymbol{\lambda}$ a generalized increment vector of order k. We discuss different models for GCF in B.

3.1 The IK metamodel

In this subsection we introduce *intrinsic Kriging* (IK) based on an IRF-k. The goal is to derive a new metamodel that uses only allowable linear combinations of order k, since they have finite variance. Let $\mathscr{M}(\mathbf{x})$ be an IRF-k with mean zero and generalized covariance function K. Then the IK metamodel is

$$\mathscr{Y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathscr{M}(\mathbf{x}). \tag{3}$$

The IK metamodel predicts $\mathscr{Y}$ at a new point $\mathbf{x}_0$ using a linear combination of observed data $\mathscr{Y}(\mathbf{x})$. With proper constraints on the weights, we guarantee that the prediction error $\hat{\mathscr{Y}}(\mathbf{x}_0) - \mathscr{Y}(\mathbf{x}_0)$ is an allowable linear combination of order k.

Cressie (1991, pp. 299–306) derives a linear predictor for the IRF-k metamodel defined in (3) with generalized covariance function K. We have the old outputs $\boldsymbol{\mathscr{Y}} = (\mathscr{Y}(\mathbf{x}_1), \ldots, \mathscr{Y}(\mathbf{x}_m))^\top$ with the generalized covariance matrix $\mathbf{K}$. The optimal linear prediction of $\mathscr{Y}$ at a new location $\mathbf{x}_0$ follows from minimizing the mean squared prediction error (MSPE) of the linear predictor:

$$\min_{\boldsymbol{\lambda}} E\left(\hat{\mathscr{Y}}(\mathbf{x}_0) - \mathscr{Y}(\mathbf{x}_0)\right)^2 \text{ such that } \hat{\mathscr{Y}}(\mathbf{x}_0) = \boldsymbol{\lambda}^\top \boldsymbol{\mathscr{Y}}. \tag{4}$$

IK should meet the condition

$$\boldsymbol{\lambda}^\top \mathbf{F} = \left(f_0(\mathbf{x}_0), \ldots, f_p(\mathbf{x}_0)\right). \tag{5}$$

This condition guarantees that the coefficients of the prediction error $\lambda_1 \mathscr{Y}(\mathbf{x}_1) + \ldots + \lambda_m \mathscr{Y}(\mathbf{x}_m) - \mathscr{Y}(\mathbf{x}_0)$ create a generalized-increment vector $\boldsymbol{\lambda}_{m+1}^\top = \left(\boldsymbol{\lambda}^\top, \lambda_0\right)$ with $\lambda_0 = -1$. This gives the variance of the IK predictor, denoted by σ^2_{IK}:

$$\sigma^2_{\text{IK}} = \text{var}(\boldsymbol{\lambda}_{m+1}^\top \boldsymbol{\mathscr{Y}}) = \sum_{i=0}^{m} \sum_{i'=0}^{m} \lambda_i \lambda_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}). \tag{6}$$

Temporarily we assume that K is known, so the optimal linear predictor is obtained through minimization of (6) subject to (5). Hence, the IK predictor is given by (4) with

$$\boldsymbol{\lambda}^\top = \left(\mathbf{K}(\mathbf{x}_0, \cdot) + \mathbf{F}\left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1}\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}(\mathbf{x}_0, \cdot)\right)\right)^\top \mathbf{K}^{-1} \tag{7}$$

where $\mathbf{K}(\mathbf{x}_0, \cdot) = (\mathbf{K}(\mathbf{x}_0, \mathbf{x}_1), \ldots, \mathbf{K}(\mathbf{x}_0, \mathbf{x}_m))^\top$ and $\mathbf{K}$ is an $m \times m$ matrix with (i, i') element $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_{i'})$. The resulting σ^2_{IK} is given by

$$\begin{aligned} &\text{MSPE}(\hat{\mathscr{Y}}(\mathbf{x}_0)) = \mathbf{K}(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{K}(\mathbf{x}_0, \cdot)^\top \mathbf{K}^{-1} \mathbf{K}(\mathbf{x}_0, \cdot) + \\ &\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}(\mathbf{x}_0, \cdot)\right)^\top \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}(\mathbf{x}_0, \cdot)\right). \end{aligned} \tag{8}$$

In practice, K is unknown so we estimate the covariance function parameters (say) $\boldsymbol{\theta}$ in $K = K(\boldsymbol{\theta})$. For this estimation we use *restricted maximum likelihood*

(REML) estimator, which maximizes the likelihood of *transformed* data that do not contain the unknown parameters of the drift. This transformation is close to the concept of *ordinary increments*. So we assume $\mathscr{Y}$ is a *Gaussian* IRF-k. The REML estimator of $\boldsymbol{\theta}$ is then found through minimization of the negative log-likelihood function

$$\ell(\boldsymbol{\theta}) = (m-q)/2\log(2\pi) - \frac{1}{2}\log\left|\mathbf{F}^\top\mathbf{F}\right| + \frac{1}{2}\log|\mathbf{K}(\boldsymbol{\theta})| + \frac{1}{2}\log\left|\mathbf{F}^\top\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{F}\right| + \frac{1}{2}\boldsymbol{\mathscr{Y}}^\top\boldsymbol{\Xi}(\boldsymbol{\theta})\boldsymbol{\mathscr{Y}} \tag{9}$$

where $q = \text{rank}(\mathbf{F})$ and $\boldsymbol{\Xi}(\boldsymbol{\theta}) = \mathbf{K}(\boldsymbol{\theta})^{-1} - \mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{F}\left(\mathbf{F}^\top\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}(\boldsymbol{\theta})^{-1}$. Finally, we replace $\mathbf{K}$ by $\mathbf{K}(\hat{\boldsymbol{\theta}})$ in (7)—to obtain $\widehat{\boldsymbol{\lambda}}$—and in (8)—to obtain $\hat{\sigma}^2_{\text{IK}}$.

We could require REML to estimate the optimal (integer) $\mathbf{k}^*$, but this would make the optimization more difficult. In our methodology, the user should try different values for $\mathbf{k}$ and pick the one which gives a better fit. Developing a procedure for finding $\mathbf{k}^*$ without user intervention is a topic for future research.

4 Stochastic simulation

The *interpolating* property of IK does not make sense for random simulation, which has sampling variability or internal noise besides the external noise or spatial uncertainty created by the fitted metamodel. In Section 4.1, we extend the theory of IK to random simulation with internal noise variances that change across the input space. In Section 4.2, we discuss an experimental design with minimum integrated MSPE.

4.1 Stochastic intrinsic Kriging (SIK)

The extension of IK to internal noise with a *constant* variance has already been studied in the literature as the *nugget effect* (geostatistics) or *jitter* (machine learning). Indeed, Cressie (1991, p. 305) briefly discusses IK with a nugget effect, replacing $K(\mathbf{h})$ by $K(\mathbf{h}) + c_0\delta(\mathbf{h})$ where $c_0 \geq 0$, $\delta(\mathbf{h}) = 0$ if $\mathbf{h} > 0$, and $\delta(\mathbf{h}) = 1$ if $\mathbf{h} = 0$. Our contribution considers internal noise with heteroscedastic variances.

Our methodology is similar to the one that is used to incorporate internal noise in Kriging and is published under different names; see Opsomer et al. (1999); Ginsbourger (2009); Ankenman et al. (2010); Yin et al. (2011). We extend the IK metamodel defined in (3) incorporating the internal noise. The value of this metamodel at replication r of the random output at $\mathbf{x}$ is

$$\mathsf{Y}_r(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}) + \varepsilon_r(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^d \tag{10}$$

where $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ denotes the internal noise at point $\mathbf{x}$. We assume that the internal noise has a Gaussian distribution with mean zero and variance $\mathsf{V}(\mathbf{x})$ and that this internal noise is independent of the external noise $\mathsf{M}(\mathbf{x})$.

Our experimental design consists of pairs $(\mathbf{x}_i, n_i),\ i = 1, \ldots, m$, where n_i denotes the number of replications at $\mathbf{x}_i$. These replications enable us to compute the classic unbiased estimators of the mean output and the internal variance:

$$\overline{\mathsf{Y}}(\mathbf{x}_i) = \frac{\sum_{r=1}^{n_i} \mathsf{Y}_{i;r}}{n_i} \text{ and } s^2(\mathbf{x}_i) = \frac{\sum_{r=1}^{n_i} \left(\mathsf{Y}_{i;r} - \overline{\mathsf{Y}}(\mathbf{x}_i)\right)^2}{n_i - 1}. \tag{11}$$

We rewrite (10) as

$$\overline{\mathsf{Y}}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}) + \overline{\varepsilon}(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^d \tag{12}$$

where $\mathsf{M}(\mathbf{x})$ is an IRF-k.

Because we assumed that $\mathsf{M}(\mathsf{x})$ and $\varepsilon(\mathbf{x})$ in (10) are independent, the *stochastic intrinsic Kriging* (SIK) predictor and its MSPE can be derived similarly to the IK in (4) and (8)—except that $\mathbf{K}_\mathsf{M}$ is replaced by $\mathbf{K} = \mathbf{K}_\mathsf{M} + \mathbf{K}_{\overline{\varepsilon}}$ where $\mathbf{K}_{\overline{\varepsilon}}$ is a diagonal matrix (so we assume that no common random numbers are used in the random simulation) with the variances of the internal noise $\mathsf{V}(\mathbf{x}_i)/n_i$ on the main diagonal; $\mathbf{K}_\mathsf{M}$ still denotes the generalized covariance matrix of IK without internal noise. We also replace $\mathscr{Y}$ in (4) and (8) by $\overline{\mathbf{Y}} = (\ \overline{\mathsf{Y}}(\mathbf{x}_1), \ldots, \overline{\mathsf{Y}}(\mathbf{x}_m)\ )^\top$. So the SIK predictor is

$$\begin{aligned}
&\widehat{\mathsf{Y}}(\mathbf{x}_0) = \boldsymbol{\lambda}^\top \overline{\mathbf{Y}} \text{ where} \\
&\boldsymbol{\lambda}^\top = \left(\mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot) + \mathbf{F}\left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1}\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\right)\right)^\top \mathbf{K}^{-1}
\end{aligned} \tag{13}$$

and its MSPE is

$$\begin{aligned}
&\mathrm{MSPE}(\widehat{\mathsf{Y}}(\mathbf{x}_0)) = \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot) + \\
&\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\right)^\top \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\right).
\end{aligned} \tag{14}$$

Note that in our experiments we estimate the MSPE from the known mean output of our test functions; see (17).

We use REML (see Section 3.1) to estimate $\boldsymbol{\theta}$, and replace $\mathbf{K}_\mathsf{M}$ by $\mathbf{K}_\mathsf{M}(\widehat{\boldsymbol{\theta}})$. We also need to estimate the internal noise V. Inspired by Ankenman et al. (2010) we use an IK metamodel for the internal noise.

$$\mathsf{V}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\sigma} + \mathsf{Z}(\mathbf{x})$$

where Z is an IRF-k independent of M. We know that $\mathsf{V}(\mathbf{x})$ is not observable, even at old points $\mathbf{x}_i$ $(i = 1, \ldots, m)$, so we let $s^2(\mathbf{x}_i)$ defined in (11) replace $\mathsf{V}(\mathbf{x}_i)$. We use IK to model the internal noise, so we assume the $s^2(\mathbf{x}_i)$ have no noise and $\widehat{\mathsf{V}}(\mathbf{x}_i) = s^2(\mathbf{x}_i)$. We replace $\mathbf{K}_{\overline{\varepsilon}}$ by $\widehat{\mathbf{K}}_{\overline{\varepsilon}} = \left(\widehat{\mathsf{V}}(\mathbf{x}_1)/n_1, \ldots, \widehat{\mathsf{V}}(\mathbf{x}_m)/n_m\right)$. Finally, we replace $\mathbf{K} = \mathbf{K}_\mathsf{M} + \mathbf{K}_{\overline{\varepsilon}}$ by $\widehat{\mathbf{K}} = \mathbf{K}_\mathsf{M}(\widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{K}}_{\overline{\varepsilon}}$ in (13) and (14). In the next subsection we explain how we choose the number of replications at each point n_i.

4.2 Design of experiments with stochastic simulation models

Like Ankenman et al. (2010) we are interested in an experimental design with low integrated MSPE (IMSPE); actually, IMSPE was introduced in Sacks et al. (1989); however, our approach is slightly different. Ankenman et al. (2010) use MSPE for the case of simple Kriging where the trend is known, whereas we use MSPE for the general case of UK where the trend is unknown.

In our design we have to allocate N replications among m old points $\mathbf{x}_i$ such that this design minimizes the IMSPE. Let $\mathscr{X}$ be the design space. Then our goal is

$$\min_{\mathbf{n}} \text{IMSPE}(\mathbf{n}) = \min_{\mathbf{n}} \int_{\mathbf{x}_0 \in \mathscr{X}} \text{MSPE}(\mathbf{x}_0, \mathbf{n}) d\mathbf{x}_0 \tag{15}$$

subject to $\mathbf{n}^\top \mathbf{1}_m \leq N$, and $\mathbf{n} = (n_1, \ldots, n_m)^\top$ where $n_i \in \mathbb{N}$.

We formulate IMSPE as

$$\begin{aligned}
&\text{IMSE}(\mathbf{n}) = \int \mathsf{K}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) d\mathbf{x}_0 - \\
&\text{trace}\left\{ \underbrace{\begin{bmatrix} \mathbf{O} & \mathbf{F}^\top \\ \mathbf{F} & \mathsf{K}(\mathbf{n}) \end{bmatrix}^{-1}}_{S(\mathbf{n})^{-1}} \int \underbrace{\begin{bmatrix} \mathbf{f}(\mathbf{x}_0)\mathbf{f}(\mathbf{x}_0)^\top & \mathbf{f}(\mathbf{x}_0)\mathsf{K}_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \\ \mathsf{K}_{\mathsf{M}}(\mathbf{x}_0, \cdot)\mathbf{f}(\mathbf{x}_0)^\top & \mathsf{K}_{\mathsf{M}}(\mathbf{x}_0, \cdot)\mathsf{K}_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \end{bmatrix}}_{\boldsymbol{\Gamma}(\mathbf{x}_0)} d\mathbf{x}_0 \right\} \\
&= \underbrace{\int \mathsf{K}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) d\mathbf{x}_0}_{\kappa} - \text{trace}\left\{ S(\mathbf{n})^{-1} \underbrace{\int \boldsymbol{\Gamma}(\mathbf{x}_0) d\mathbf{x}_0}_{\mathbf{W}} \right\} \\
&= \kappa - \text{trace}\left\{ \mathbf{S}(\mathbf{n})^{-1} \mathbf{W} \right\} = \kappa - \sum_{q,q'=1}^{p+1+m} [\mathbf{S}(\mathbf{n})^{-1}]_{qq'} W_{qq'} \\
&= \kappa - \mathbf{1}^\top \left[\mathbf{S}(\mathbf{n})^{-1} \circ \mathbf{W} \right] \mathbf{1}
\end{aligned}$$

where $\circ$ is the Hadamard product.

Then for the minimization problem we write the Lagrangian function

$$L(\mathbf{n}) = \text{IMSPE}(\mathbf{n}) + \eta(N - \mathbf{1}^\top \mathbf{n}).$$

The m first-order optimality conditions are

$$\frac{\partial L(\mathbf{n})}{\partial n_i} = \frac{\partial \text{IMSPE}(\mathbf{n})}{\partial n_i} + \eta = 0, \qquad (i = 1, \ldots, m).$$

Application of linear algebra gives

$$\begin{aligned}
\frac{\partial \text{IMSPE}(\mathbf{n})}{\partial n_i} &= -\mathbf{1}^\top \left[\mathbf{W} \circ \frac{\partial \mathbf{S}(\mathbf{n})^{-1}}{\partial n_i} \right] \mathbf{1} \\
&= -\frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} \mathbf{1}^\top \left[\mathbf{W} \circ \left(\mathbf{S}(\mathbf{n})^{-1} \mathbf{J}^{(ii)} \mathbf{S}(\mathbf{n})^{-1} \right) \right] \mathbf{1} \\
&= -\frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} \text{trace} \left(\mathbf{W}^\top \mathbf{S}(\mathbf{n})^{-1} \mathbf{J}^{(ii)} \mathbf{S}(\mathbf{n})^{-1} \right) \\
&= -\frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} \text{trace} \left(\mathbf{J}^{(ii)} \mathbf{S}(\mathbf{n})^{-1} \mathbf{W} \mathbf{S}(\mathbf{n})^{-1} \right) \\
&= -\frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} \sum_{q,q'=1}^{p+1+m} \left[\mathbf{J}^{(ii)} \right]_{qq'} \left[\mathbf{S}(\mathbf{n})^{-1} \mathbf{W} \mathbf{S}(\mathbf{n})^{-1} \right]_{qq'} \\
&= -\frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} \left[\mathbf{S}(\mathbf{n})^{-1} \mathbf{W} \mathbf{S}(\mathbf{n})^{-1} \right]_{p+1+i,p+1+i}
\end{aligned}$$

where $\mathbf{J}^{(ii)}$ is a $(p+1+m)\times(p+1+m)$ matrix with 1 in position $(p+1+i, p+1+i)$ and 0 elsewhere.

When we try to compute n_i^*, which denotes the optimal allocation of the total number of replications N over the m old points. We run into the problem that n_i^* is determined by $\mathbf{K}_M$ (external noise) and $\mathbf{K}_{\bar{\varepsilon}}$ (internal noise). To solve this problem, we follow Ankenman et al. (2010) and ignore $\mathbf{K}_{\bar{\varepsilon}}$ so $\mathbf{K} \approx \mathbf{K}_M$; moreover, we relax the integrality condition. Altogether we derive

$$n_i^* \approx N \frac{\sqrt{\mathsf{V}(\mathbf{x}_i) C_i}}{\sum_{i=1}^m \sqrt{\mathsf{V}(\mathbf{x}_i) C_i}} \text{ with } C_i = \left[\mathbf{S}^{-1} \mathbf{W} \mathbf{S}^{-1} \right]_{p+1+i,p+1+i}, \tag{16}$$

so both the internal noise and the external noise affect the allocation.

Because we need to estimate $\mathbf{K}_\mathsf{M}$ and $\mathbf{K}_{\bar{\varepsilon}}$, we use the *two-stage approach* proposed in Ankenman et al. (2010, p. 378). In Stage 1, we obtain a pilot sample of $m_1 < m$ points and allocate $n_0 > 10$ replications to each point. This enables us to estimate $\mathbf{K}_\mathsf{M}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\mathsf{V}}$. In Stage 2, we first select $m - m_1$ input combinations jointly, and then we optimally allocate the $N - m_1 n_0$ additional replications over the m old input combinations using (16). In the next section we discuss the application of this approach to the M/M/1 example.

5 Numerical experiments with IK and SIK

In this section we present our numerical experiments with deterministic simulations and random simulations analyzed by IK and SIK, respectively; moreover, we compare IK and SIK with classic Kriging. In these experiments we use a zero degree polynomial for the trend ($p = 0$), so UK becomes OK. In deterministic simulation we study the performance of OK versus IK. In random simulation we study the random metamodels developed based on OK and IK which account for internal noise; we call them SK (stochastic Kriging) and SIK.

We tried to use the MATLAB code developed by Ankenman et al. (2010) and Yin et al. (2011) to experiment with these Kriging variants (OK for deterministic simulation and SK for random simulation), but on our PC these MATLAB codes crashed in experiments with $d > 1$. So we use the R package `mlegp` to implement OK and SK; see Dancik (2013) for more details on `mlegp`. To save time, we implemented our MATLAB code only for IK and SIK (not for OK and SK).

We run some initial experiments to select the 'best' covariance function among valid alternatives where 'best' means lowest MSE. In IK and SIK, the best candidate was the *integrated Brownian motion* covariance function defined in (20).

We study several values for the number of old points m; namely, all integers between $10d - 5$ and $10d + 5$ where $m = 10d$ is proposed in Loeppky et al. (2009).

We experiment with several values for d (input dimensionality). The simplest case is $d = 1$, which may give valuable intuitive insight. In practice, however, $d > 1$ so we also study test functions with such d values.

To evaluate the performance of OK versus IK and SK versus SIK, we select $m_0 = 100d$ *new* points $\mathbf{x}_0$. For $d = 1$ we select m and m_0 equispaced points; for $d > 1$ we use *Latin hypercube sampling* (LHS) to select m and m_0 space-filling points. For LHS we use the MATLAB function `lhsdesign`. We quantify the performance of different metamodels through the RMSE for the m_0 new points:

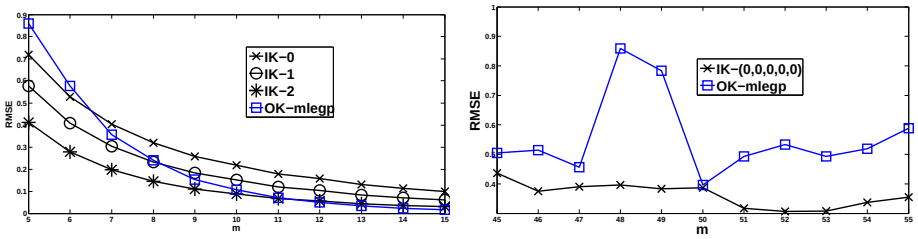
$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{m_0}(Y_t - \widehat{Y}_t)^2}{m_0}}. \tag{17}$$

5.1 Deterministic simulation experiments

In this subsection we present the results of our experiments with deterministic test functions with $d = 1, 2, 3, 5$.

We start with a *monotonic* function with $d = 1$; namely, $f(x) = 1/[x(x-1)]$ and $1 < x \leq 2$, inspired by the mean steady-state waiting-time as a function of the service rate x with the arrival rate equal to 1 in a single-server queue with Markovian arrival and service processes (denoted by M/M/1). This function increases monotonically, and increases drastically as $x \downarrow 1$. For this function, Figure 1a shows the RMSE of IK-k and OK versus different values for m. We observe that IK with $k = 2$ performs better than OK for $m \leq 11$; for larger m the difference between IK-2 and OK becomes small.

Next we experiment with several functions that are popular in optimization; see Dixon and Szego (1978) and `http://www.sfu.ca/~ssurjano/index.html`. Actually, we experiment with (i) Six-hump camel-back with $d = 2$ (ii) Hartmann-3 with $d = 3$ (iii) Levy-3 with $d = 3$ (iv) Ackley-5 with $d = 5$. We discuss the Ackley-5 function in detail, and define the first three functions in C.



(a) Monotonic one-dimensional function

(b) Ackley-5 function

Figure 1: RMSE versus m with IK-$\mathbf{k}$ and OK

The Ackley-5 function is defined as

$$f(\mathbf{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{5}\sum_{i=1}^{5}x_i^2}\right) - \exp\left(\frac{1}{5}\sum_{i=1}^{5}\cos(2\pi x_i)\right) + 20 + \exp(1),$$

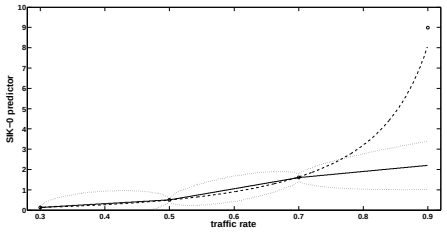
with $-2 \leq x_i \leq 2,\ i = 1, \ldots, 5$. For experiments with this function, Figure 1b shows that IK-$(0,0,0,0,0)$ gives the lowest RMSE for all values of m. We also experiment with different $\mathbf{k}$, and find that the RMSE deteriorates compared with $\mathbf{k} = (0,0,0,0,0)^\top$.

The results for the three other test functions show that IK performs better than OK except for the Hartmann-3 function; to be the winner, IK needs an appropriate value for the parameter $\mathbf{k}$ (which is not always $\mathbf{k} = \mathbf{0}$). The figures for the experiments with these functions can be found in Figure 5.

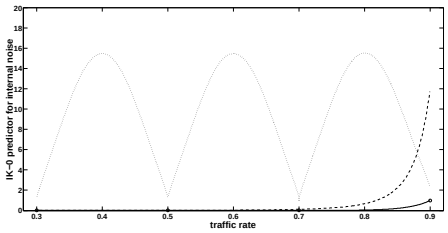
5.2 Random simulation experiments

In this subsection, we first discuss the two-stage approach detailed in Section 4.2 for the M/M/1 model (defined in Section 5.1) to show how to use SIK in real applications. Then we compare SIK and SK for different test function simulations. However, we do not use the two-stage approach to compute the optimal number of replications at each point. Instead, we select n_i proportional to the *true* value of the variance of the internal noise at each point. The reason for this approach is that we focus on comparing SIK and SK, and not on the selection of the number of replications.

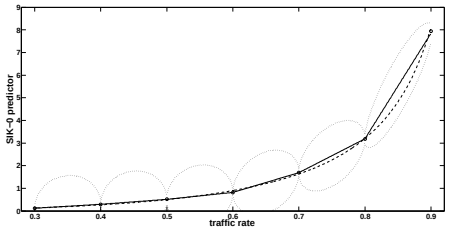
We design our M/M/1 experiment as follows. We fix the arrival rate at 1, and vary the service rate. This gives the traffic rate $\rho = 1/x$. To tackle the problem of selecting an initial state (at $T = 0$) and a warm-up period, we select the initial number of customers in the system equal to the mean steady-state number $\rho/(1-\rho)$ and we collect waiting times starting at $T = 0$. In each replication we select as terminating event $T = 3,000$ (runlength); i.e., we calculate the average waiting time of customers in the system from time 0 to 3,000. So, for all x we control the internal noise through the number of replications.



(a) Stage 1



(b) IK-0 metamodel for the internal noise



(c) Stage 2

Figure 2: Two-stage approach for the M/M/1 queue; the dotted curve denotes the true function, and the solid curve denotes the metamodel

In Stage 1, we select $\rho = 0.3, 0.5, 0.7, 0.9$ and obtain 20 replications of length T at each point. Then we fit a SIK-0 metamodel to the simulation I/O data. Figure 2a shows that the average waiting time in the queue metamodel from Stage 1. The circles represent the simulated data, the solid curve represents the SIK-0 metamodel, surrounded by $\pm\sqrt{\text{MSPE}(\widehat{\mathsf{Y}}(x_0))}$. The dashed curve represents the true function. SIK is not an interpolator, so it does not pass through the data points, as $x = 0.8$ clearly shows. Note that the $\pm\sqrt{\text{MSPE}(\widehat{\mathsf{Y}}(x_0))}$ intervals which account for internal and external noise cover the true function except for the region between $x = 0.7$ and $x = 0.9$. Figure 2b shows the IK-0 metamodel for the internal noise $\mathsf{V}(x_0)$. Here the metamodel is an interpolator and goes through the data points; clearly, the simulation gives a poor estimate of $\mathsf{V}(0.9)$.

In Stage 2, we use the information obtained in Stage 1 to apply (16) and optimally allocate $N = 500$ replications over the four old points and three new points $x = 0.4, 0.6, 0.8$. The metamodel for the internal noise helps us for these three new points, which were not simulated in Stage 1. The estimated optimal allocation is $n^* = 1, 1, 2, 3, 12, 126, 355$. In some points we have already simulated more replications than is optimal, and in other points we have to simulate additional replications. We fit a SIK-0 metamodel to data obtained from Stage 2. Figure 2c shows the SIK-0 metamodel for the data obtained from Stage 2. This figure shows that the new metamodel is close to the true function, and the MSPE intervals cover the true function. This figure demonstrates the classic shape of the MSPE intervals; namely, their length is close to zero at the

old points, and increases away from these points; however, Figure 4 in Ankenman et al. (2010) shows "nearly constant" length.

We continue this section with the comparison of SIK and SK for our test functions. In all these functions we select the number of replications at each point proportional to the true value of internal noise. In the M/M/1 example, the internal noise function is $\mathsf{V}(x_i)/T \approx 4/(Tx_i(1-1/x_i)^4)$; see Whitt (1989, p. 1350). So $n_i = \lfloor \mathsf{V}(x_i)/\sum_{i=1}^{m} \mathsf{V}(x_i) \rfloor B$ where B is the total number of replications. In the other test functions with higher dimensionality we augment the deterministic response with heteroscedastic noise; namely, $\mathsf{V}(\mathbf{x}_i) = (1 + |y(\mathbf{x}_i)|)^2$, like Wan et al. (2010) does in an experiment with linear regression metamodels.

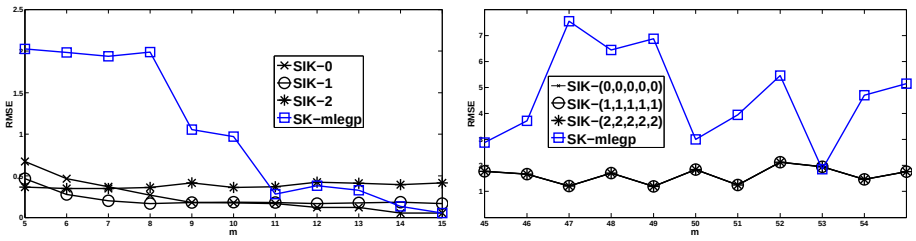
Figure 3a shows the RMSE of SIK-k and SK versus m for the M/M/1 simulation. SIK-0 gives smaller RMSEs for all m. Figure 3b shows the results for all three $\mathbf{k}$ values for the Ackley-5 simulation. We also remark that changing k from 0 to 1 or 2 in all coordinates give almost identical results. In other test functions SIK-$\mathbf{0}$ gives smaller RMSE in almost all m points except in the Levy-3 function where SIK and SK give almost identical RMSEs for all m. The figures for these test function can be found in Figure 5.

To explain these results, we may use the variogram $\gamma(h)$. We defined $\gamma(h)$ in Section 2. The plot of $\gamma(h)$ versus h indicates whether the process is *stationary* ($\gamma(h)$ stabilizes for large h) or *non-stationary* ($\gamma(h)$ increases with h). The shape of the variogram plot helps us to understand the behavior of $Y(\mathbf{x})$ and its increments of order k. We define and give the plots for the *empirical* variogram $\hat{\gamma}(h)$ in A, which we discuss next.

The panels in Figure 4 show $\hat{\gamma}(h)$ for three test functions; namely, *deterministic* monotonic one-dimensional in panel (a), *deterministic* camel-back in panel (b), and *noisy* Levy-3 test functions in panel (c). Figure 4a (from left to right) shows $\hat{\gamma}(h)$ for $Y(\mathbf{x})$ and its increments of order 0, 1, and 2. We observe that the variogram for $Y(\mathbf{x})$ shows strong non-stationarity and the variogram for its higher-order increments gets closer to stationarity. This observation explains the better RMSE results for IK in Figure 1a. Figure 4b (from left to right) shows $\hat{\gamma}(h)$ for $Y(\mathbf{x})$ and its increments of order 0 and 1. We observe that the variogram for $Y(\mathbf{x})$ shows non-stationarity behavior and the variogram for its *first* order increments gets closer to stationarity. This observation explains the better RMSE results for IK-$(1,1)$ in Figure 5a. Figure 4c (from left to right) shows $\hat{\gamma}(h)$ for $Y(\mathbf{x})$ and its increments of order 0. We observe that the two variograms have the same pattern. This observation explains the *small* difference between RMSE results for SIK and SK in Figure 5f.

6 Conclusions

Using IRFs, we derived IK for deterministic simulation, and SIK for random simulation. Additionally, we derived a two-stage approach for sample size allocation in SIK for random simulation. Next we numerically compared the performance—measured by RMSE—of IK with OK and SIK with SK. The main



(a) M/M/1 queue (b) Ackley-5

Figure 3: RMSE versus m with SIK and SK

conclusion is that in most experiments IK and SIK give smaller RMSEs than OK and SK.

In any Kriging model—including OK, SK, IK, and SIK—we must choose a specific type of covariance function. The best choice varies with the type of Kriging metamodel and the type of simulation model. We chose covariance functions that seemed best for the specific Kriging model. Obviously, finding guidelines for the choice of the appropriate covariance function requires more research.

Acknowledgement

We thank Barry L. Nelson (Northwestern University) for his MATLAB code for the M/M/1 queue simulation model, and Peter Salemi (Northwestern University) for sharing his R code for the integrated Brownian covariance functions.

A Empirical variogram plots

In Figure 4, we plot the *empirical* variogram $\hat{\gamma}(h)$ defined in (18) versus the Euclidean norm $h = \|\mathbf{x}_i - \mathbf{x}_{i'}\|$.

$$\hat{\gamma}(h) = \frac{1}{2\,|N_h|} \sum_{(\mathbf{x}_i, \mathbf{x}_{i'}) \in N_h} \left[Y(\mathbf{x}_i) - Y(\mathbf{x}_{i'})\right]^2 \tag{18}$$

where N_h is the set of pairs of observations and $|N_h|$ is the number of pairs in the set.

B Models for generalized covariance functions (GCF)

We begin with *isotropic polynomial* covariance functions $K(h = \|\mathbf{h}\|)$ where h is the Euclidean norm; by definition, isotropic covariance functions depend only

on h. Matheron (1973) developed the following $K(h = \|\mathbf{h}\|)$ for an IRF-k:

$$K(h) = \begin{cases} -\theta_1 h, & k = 0 \\ -\theta_1 h + \theta_2 h^3, & k = 1 \\ -\theta_1 h + \theta_2 h^3 - \theta_3 h^5, & k = 2 \end{cases}$$

where for $k = 0$, 1, or 2 we have the constraints $\theta_1 \geq 0$, $\theta_3 \geq 0$, and $\theta_2 \geq [(20/3)(1 + (2/(d+1))\theta_1\theta_3]^{1/2}$ in $\mathbb{R}^d$; obviously, for $k = 0$ we have $\theta_2 = 0$ and $\theta_3 = 0$ and for $k = 1$ we have $\theta_3 = 0$. For general k, the isotropic polynomial generalized covariance function is

$$K(h) = \sum_{l=0}^{k} (-1)^{l+1} \theta_{l+1} h^{2l+1} \text{ with } h = \|\mathbf{h}\| \geq 0$$

where $\theta_1, \ldots, \theta_{l+1}$ must satisfy

$$\sum_{l=0}^{k} \frac{\theta_{l+1} \Gamma\left((2l+1+d)/2\right)}{\pi^{2l+2+(d/2)} \Gamma\left(1 + (1/2)(2l+1)\right)} \rho^{-d-2l+1} \geq 0 \text{ for any } \rho \geq 0$$

where $\Gamma(\cdot)$ denotes the Gamma function.

We use the first two examples of IRF-k in Section 3, to introduce new generalized covariance functions.

1. The $(k+1)$ integral of a zero-mean stationary random function with covariance $C(h = x - x')$ is an IRF-k with generalized covariance function

$$K(h) = (-1)^{k+1} \int_0^h \frac{(h-u)^{2k+1}}{(2k+1)!} C(u) du.$$

There are many choices for C. The common choices are exponential $C(h) = \exp(-\theta h)$, Gaussian $C(h) = \exp(-\theta h^2)$, or Matern. Now we derive the generalized covariance functions for $k = 0$ in two cases of $C(\cdot)$.

(a) Gaussian $C(\cdot)$:

$$\begin{aligned} K(h) &= (-1) \int_0^h (h-u) \exp(-\theta u^2) du \\ &= \left(1 - \exp(-\theta h^2) - h\sqrt{\pi}\,\mathrm{erf}(\sqrt{\theta} h)\right) / 2\theta \end{aligned}$$

where $\mathrm{erf}(\cdot)$ is the error function.

(b) Exponential $C(\cdot)$:

$$\begin{aligned} K(h) &= (-1) \int_0^h (h-u) \exp(-\theta u) du \\ &= \frac{h}{\theta} \left(\exp(-\theta h) - 1\right) + \frac{1}{\theta^2} \left(1 - (1 + \theta h) \exp(-\theta h)\right). \end{aligned}$$

2. Integrating k times a Brownian motion $B(x)$ with covariance function $C(x, x'; \theta) = \theta \min(x, x')$ with $x, x' \in [0, 1]$ and $\theta \geq 0$ gives an IRF-k with generalized covariance function

$$K(x, x'; \theta) = \theta \int_0^1 \frac{(x-u)_+^k (x'-u)_+^k}{(k!)^2} du \tag{19}$$

where $\theta \geq 0$; see Berlinet and Thomas-Agnan (2004, p. 92). Furthermore, Salemi et al. (2013, pp. 546–547) suggests to add polynomial terms to this covariance to avoid $B(x)$ becoming zero at $x = 0$. In our experiments we add a constant term $\theta_0 \geq 0$ to (19). Note that the Brownian motion is a special case of the fractional Brownian motion B_H, which is an IRF-0 with the covariance function

$$\operatorname{cov}\left(B_H(x), B_H(x')\right) = \frac{1}{2}\left(x^{2H} + x'^{2H} - (x - x')^{2H}\right), \qquad 0 < H < 1$$

where $H = 1/2$ gives the usual Brownian motion.

Besides the preceding isotropic functions, we are also interested in *anisotropic* generalized covariance functions. We use the same idea; i.e., the multiplication of valid covariance functions gives a valid covariance function, so we multiply the covariance functions per input dimension. So the anisotropic version of K for the covariance function of the integrated Brownian motion defined in (19) is

$$K(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \prod_{g=1}^{d} \left(\theta_{0;g} + \theta_{1;g} \int_0^1 \frac{(x_g - u_g)_+^{k_g} (x'_g - u_g)_+^{k_g}}{(k_g!)^2} du_g \right) \tag{20}$$

where $\boldsymbol{\theta} = (\theta_{0;1}, \theta_{1;1}, \theta_{0;2}, \ldots, \theta_{0;d}, \theta_{1;d}) \geq 0$.

The anisotropic covariance function accepts different k values for different input dimensions, so we have a vector $\mathbf{k} = (k_1, \ldots, k_d)^\top$ instead of a single scalar k. Anisotropic covariance functions handle each input dimension separately, so they are more flexible than isotropic covariance functions. However, this comes at the cost of estimating more parameters.

C Test functions with dimensionality $d > 1$

In this appendix we define the test functions with $d > 1$, and give results of our experiments.

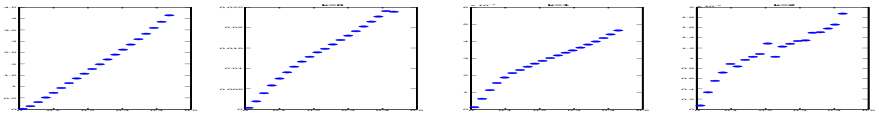
1. Six-hump camel-back with $-2 \leq x_1 \leq 2,\ -1 \leq x_2 \leq 1$

$$f(x_1, x_2) = 4x_1^2 - 2.1x_1^4 + x_1^6/3 + x_1 x_2 - 4x_2^2 + 4x_2^4$$

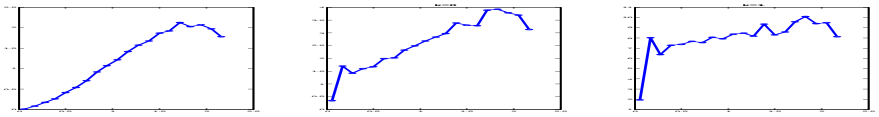
2. Hartmann-3 function with $0 \leq x_i \leq 1,\ i = 1, 2, 3$

$$f(x_1, x_2, x_3) = -\sum_{i=1}^{4} \alpha_i \exp[-\sum_{j=1}^{3} A_{ij}(x_j - P_{ij})^2]$$

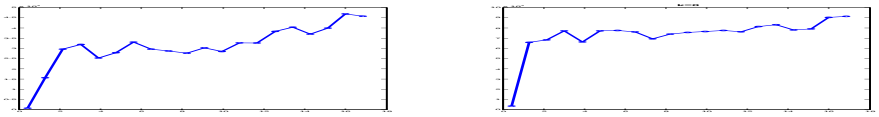
with $\boldsymbol{\alpha} = (1.0, 1.2, 3.0, 3.2)^\top$ and A_{ij} and P_{ij} given in Table 1.



(a) Monotonic function: $\hat{\gamma}(h)$ for $Y(\mathbf{x})$ and its increments of order 0, 1, and 2



(b) Camel-back function: $\hat{\gamma}(h)$ for $Y(\mathbf{x})$ and its increments of order 0 and 1

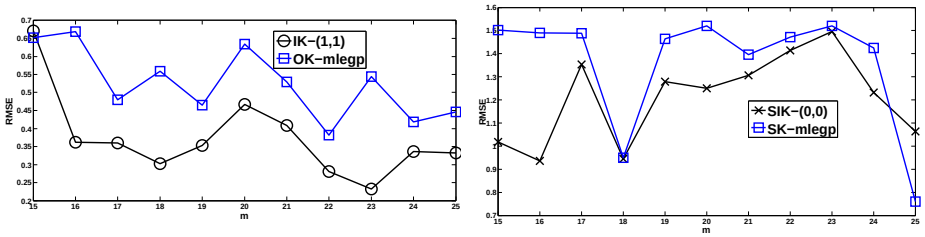


(c) Levy-3 function: $\hat{\gamma}(h)$ for $Y(\mathbf{x})$ and its increments of order 0

Figure 4: Empirical variogram $\hat{\gamma}(h)$ versus h

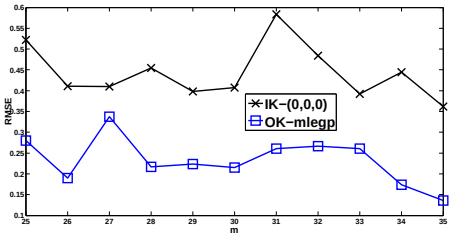
Table 1: Parameters A_{ij} and P_{ij} of the Hartmann-3 function

A_{ij}			P_{ij}		
3	10	30	0.36890	0.1170	0.26730
0.1	10	35	0.46990	0.43870	0.74700
3	10	30	0.10910	0.87320	0.55470
0.1	10	35	0.03815	0.57430	0.88280

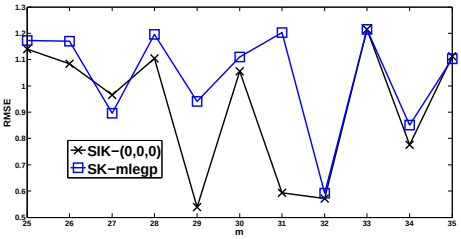


(a) IK-**k** and OK in camel-back function

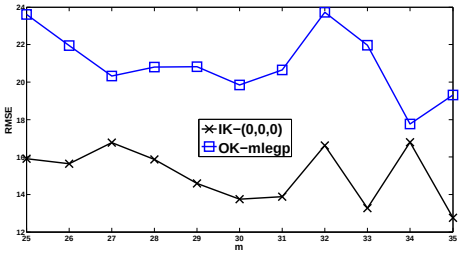
(b) SIK-**k** and SK in camel-back function



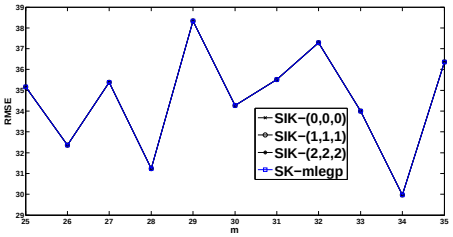
(c) IK-**k** and OK in Hartmann-3 function



(d) SIK-**k** and SK in Hartmann-3 function



(e) IK-**k** and OK in Levy-3 function



(f) SIK-**k** and SK in Levy-3 function

Figure 5: RMSE versus m in deterministic simulations (left panels) and random simulations (right panels)

3. Levy-3 function with $-10 \leq x_i \leq 10,\ i = 1, 2, 3$

   $$f(\mathbf{x}) = \sin^2(\pi w_1) + (w_2 - 1)^2[1 + 10\sin^2(\pi w_2 + 1)] + (w_3 - 1)^2[1 + \sin^2(2\pi w_3)]$$

   where $w_i = 1 + (x_i - 1)/4$.

References

Ankenman, B., Nelson, B., and Staum, J. (2010). Stochastic Kriging for simulation metamodeling. *Operations Research*, 58:371–382.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

Chilès, J. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley.

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

Dancik, G. (2013). *mlegp: Maximum Likelihood Estimates of Gaussian Processes*. R package version 3.1.4.

Dixon, L. and Szego, G., editors (1978). *Towards Global Optimisation 2*. Elsevier Science Ltd.

Ginsbourger, D. (2009). *Multiples Métamodèles pour l'Approximation et l'Optimisation de Fonctions Numériques Multivariables*. PhD dissertation, Ecole Nationale Supérieure des Mines de Saint-Etienne.

Loeppky, J., Sacks, J., and Welch, W. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51:366–376.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468.

Opsomer, J., Ruppert, D., Wand, M., Holst, U., and Hossjer, O. (1999). Kriging with nonparametric variance function estimation. *Biometrics*, 55(3):704–710.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Salemi, P., Staum, J., and Nelson, B. (2013). Generalized integrated brownian fields for simulation metamodeling. In *2013 Winter Simulation Conference (WSC)*, pages 543–554.

Santner, T., Williams, B., and Notz, W. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging (Springer Series in Statistics)*. Springer.

Vazquez, E., Walter, E., and Fleury, G. (2005). Intrinsic Kriging and prior information. *Applied Stochastic Models in Business and Industry*, 21(2):215–226.

Wan, H., Ankenman, B., and Nelson, B. (2010). Improving the efficiency and efficacy of controlled sequential bifurcation for simulation factor screening. *INFORMS Journal on Computing*, 22(3):482–492.

Whitt, W. (1989). Planning queueing simulations. *Management Science*, 35(11):1341–1366.

Yin, J., Ng, S., and Ng, K. (2011). Kriging meta-model with modified nugget effect: an extension to heteroscedastic variance case. *Computers and Industrial Engineering*, 61(3):760–777.